INST414 - Data Science Techniques
Kaggle Competition

Round 1: Tue. 11/24, 11:59 pm. Round 2: Fri. 12/4, 11:59 pm. Memo: Sun. 12/6, 11:59 pm.
Due to the nature of the leaderboard, we cannot extend Round 2.

This is an exercise of text classification, through the platform of an online data science competition: `https://www.kaggle.com/c/umd-inst414-20f-imdb`.

You may use the following invitation link to join the competition but please refrain from posting the invitation link elsewhere: `https://www.kaggle.com/t/450ce9b855bb4d28a085a26b4126b712`.

This is a text classification task. Every document (a line in the data file) is a movie review from IMDB. Your goal is to classify each document into ONE of the two categories, based on whether it needs simplification: 1 if the review is positive; 0 if the review is negative.

The training data contains 10,000 reviews, already labeled with one of the above categories. The test data contains 5,000 reviews that are unlabeled. The submission should be a .csv (comma separated free text) file with a header line "Id,Category" followed by exactly 5,000 lines. In each line, there should be exactly two integers, separated by a comma. The first integer is the line ID of a test question (0 - 5,000), and the second integer is the category your classifier predicts one of (0,1).

You can make 10 submissions per day. Once you submit your results, you will get an accuracy score computed based on 50% of the test data. This score will position you somewhere on the leaderboard. Once the competition ends, you will see the final accuracy computed based on the other 50% of the test data. The evaluation metric is the accuracy of your classifier - so the higher the better. You can use any classifiers, any combination (or subset) of features.

Grading: The grading of this assignment include *three* parts: two rounds of competitions and a 1-page memo. For the two rounds of competitions, we will evaluate your classification result using the accuracy, namely the ratio of the reviews that are classified correctly.

1. Competition Round 1 (5 + 0.5 pts): The first round ends at Tue. Nov 24th 11:59PM. The evaluation is simple:
   (a) Everyone who beats a correctly implemented Naive Bayes classifier (0.82560 accuracy on the public test set) gets full points.
   (b) If your cannot beat a NB baseline, your score will be deducted based on the accuracy.
   (c) If you beats a well-tuned SVM classifier (0.88120 on the public test set), you will receive 0.5 bonus point.
2. Competition Round 2 (10 pts): The second rounds ends 10 days later, at Fri, December 4th 11:59PM. You will receive at least 8 points if and only if the accuracy score of your best classifier beats a correctly implemented SVM classifier (0.84680 on the public test set). The other 2 points will be given according to your position on the leaderboard. The formula to compute your grade:

   $$\text{grade} = 7 + 3 * 2 / \log_2(2 + \text{rank})$$

   Note that **the winner can get as much as 10.8 points!** However, if your submission did not beat the SVM baseline, your score will be less than 8 points regardless of the ranking.
3. One Page Memo (5 pts): Please submit a one page memo (in .pdf) describing the preprocessing, features, models, and parameter tuning you explored and the corresponding results. Please write down your name and the display name you used in the competition. In addition, please also submit the source code in your submissions. If you received help from anyone, you should list the name(s) in your memo.

Have fun! And don't waste your quota of submissions!